Michael Wang michaelwang 438275 Terry Lyu terrylu 435091
HW 4

1a. No, because we aren't inputting any parameters on the command line, there are no differences in job execution

1b. The difference in execution here is the command we put into the command line.

1c. hadoop jar avgjob.jar solution/AvgWordLength –D caseSensitive=false Shakespeare /newshake

The order of the command line inputs does matter, the last two have to be the input and output locations.

A = 3.276

w = 4.373

z = 5.053

2a. IP Address – Date/Time Type of Request Folder Http Version response code size

Website request log files.

5000 lines in weblog/4477843 lines in full file = .00111

2b. Mapper input:

(byte offset, 10.223.157.186)

(byte offset, 10.223.157.186)

(byte offset, 10.216.113.172)

(byte offset, 10.216.113.172)

(byte offset, 10.216.113.172)

Mapper output:

(10.223.157.186, 1)

(10.223.157.186, 1)

(10.223.157.186, 1)

(10.216.113.172, 1)

(10.216.113.172, 1)

Reducer Input:

(10.223.157.186, [1, 1, 1])

(10.216.113.172, [1, 1])

Reducer Output:

(10.223.157.186, 3)

Michael Wang michaelwang 438275 Terry Lyu terrylu 435091
HW 4

(10.216.113.172, 2)

Reducer is taking in hits from different ip address and sum them up to keep track of total number of hits from each ip address.

Part c part d . please refer to the code submitted in the repository

Part e : the command is

Hadoop jar log5.jar stubs/ProcessLogs -fs=file:/// -jt=local access_log accesstest

Part f

Suppose the five lines in part b is the testlog file,

Then there are two different ip addresses, every line results in a count

Part g

What to bear in mind? We need instead of using the local file system, we should use hdfs system,

There are 333923 lines(different ip addresses) in the complete log

There are 35 hits for 10.1.100.199

There are 1 hits for 10.1.100.5

There are 21 hits for 10.99.99.58

Why are they sorted? In a MapReduce program, the output key-value pairs from the Mapper are automatically sorted by keys